The performance of permutations and exponential random graph models when analysing animal networks

Julian C. Evans[1], David N. Fisher[2], Matthew J. Silk[3,4*]

Running header: Evaluating ERGMs and permutations for network analysis

1. Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
2. School of Biological Sciences, University of Aberdeen, King's College, Aberdeen, AB23 3FX, United Kingdom
3. Centre for Ecology and Conservation, University of Exeter Penryn Campus, Treliever Road, Penryn, Cornwall, TR10 9FE, United Kingdom
4. Environment and Sustainability Institute, University of Exeter Penryn Campus, Treliever Road, Penryn, Cornwall, TR10 9FE, United Kingdom

* corresponding author: matthewsilk@outlook.com

The performance of permutations and exponential random graph models when analysing animal networks

Running header: Evaluating ERGMs and randomisations for network analysis

Abstract

Social network analysis is a suite of approaches for exploring relational data. Two approaches commonly used to analyse animal social network data are permutation-based tests of significance and exponential random graph models. However, the performance of these approaches when analysing different types of network data has not been simultaneously evaluated. Here we test both approaches to determine their performance when analysing a range of biologically realistic simulated animal social networks. We examined the false positive and false negative error rate of an effect of a two-level explanatory variable (e.g. sex) on the number and combined strength of an individual's network connections. We measured error rates for two types of simulated data collection methods in a range of network structures, and with/without a confounding effect and missing observations. Both methods performed consistently well in networks of dyadic interactions, and worse on networks constructed using observations of individuals in groups. Exponential random graph models had a marginally lower rate of false positives than permutations in most cases. Phenotypic assortativity had a large influence on the false positive rate, and a smaller effect on the false negative rate for both methods in all network types. Aspects of within- and between-group network structure influenced error rates, but not to the same extent. In grouping-event based networks, increased sampling effort marginally decreased rates of false negatives, but increased rates of false positives for both analysis methods. These results provide guidelines for biologists analysing and interpreting their own network data using these methods.

Key words: social network analysis, permutation, randomisation, exponential random graph model

Introduction

Essentially all animals engage in some form of social interaction, ranging from interacting with large numbers of individuals while living in groups, to mating and competitive interactions among otherwise solitary organisms (Frank 2007). Social interactions are key for various aspects of organism biology, such as development (Berman and Kapsalis 1999; Bautista et al. 2015), movement and dispersal (Sumpter 2006; Strandburg-Peshkin et al. 2017), and mating (Clutton-Brock et al. 1997; Cheney et al. 2016). As such, the development of methods that quantify social interactions in a wide range of taxa and enable accurate inference of the underlying causes of variation in social connectivity is key (Krause et al. 2014).

Studying the social lives of animals can be challenging, as the nature of their associations, interactions and relationships can be difficult to observe and quantify in a manner consistent across species and contexts. In the last two decades much headway has been made by incorporating the techniques of social network analysis (SNA) into ecological and evolutionary studies (Webber and Vander Wal 2019). In a social network, individuals ("nodes") interact with others (connected by "edges") forming a network, which can be represented as a pairwise adjacency matrix. Initially developed in sociology to study human interactions (Wasserman and Faust 1994), SNA has now been widely applied to the interactions of mammals such as primates (Sade 1972), cetaceans (Lusseau 2003) and elephants (Wittemyer et al. 2005), as well as birds (Myers 1983), lizards (Leu et al. 2010), fish (Croft et al. 2004) and insects (Fewell 2003).

Social network data often violate assumptions of conventional statistical approaches through being non-independent as a result of the relational nature of the data being analysed (James et al. 2009; Croft et al. 2011). Additionally social network data can often contain biases imposed by the method of data collection (Franks et al. 2010), such as when observations are skewed towards the most detectable individuals and/or in the environments that are the easiest to make observations in. As a result, some methods of data collection can imbue even randomly generated networks with seemingly biological patterns (Franks et al. 2010). While association measures have been developed that can control for some of these biases (Whitehead and James 2015), it remains important to control for them in subsequent analyses (James et al. 2009; Croft et al. 2011). In addition, response variables obtained from social networks are frequently non-Gaussian, and often zero-inflated, which increases the complexity of the statistical modelling required (Krivitsky 2012, 2015). Finally, individuals may often be missed, or interactions not detected, which can influence both individual network metrics but also whole-network structure (Franks et al. 2010; Silk et al. 2015; Davis et al. 2018).

To deal with this challenging data analysis, a suite of methods has been developed specifically for SNA (Wasserman and Faust 1994). We focus on two common choices for analysing the social networks of animals. The first of these are permutation-based approaches (Bejder et al. 1998; Anderson et al. 1999; also referred to as "randomisation-based approaches" or simply "randomisations"). Here, the observed data (either raw data prior to constructing the network or the network itself) are permuted, with analytical outputs from the resulting randomised networks compared to equivalent outputs from the observed data to test for statistical significance. The advantages of this are twofold. First, using a permutation-based approach does not make the same assumptions about the independence or normality of model residuals as more conventional statistical approaches do. Second, by constraining the permutations in particular ways it is possible to control for biases generated by methods of data collection (Bejder et al. 1998; James et al. 2009; Croft et al. 2011), which is particularly important when social relationships are inferred from data on spatio-temporal co-occurrence or group membership (Whitehead and Dufault 1999; Franks et al. 2010).

The most basic permutation methods perform swaps on the network itself by swapping the identity of nodes or edges. However, more complex approaches permute the collected data prior to construction of the network (the "datastream"), and can offer greater ability to control for biases in data collection especially for methods which infer social relationships from group membership (Croft et al. 2008; Farine and Whitehead 2015; Farine 2017). By permuting the identities of the individuals within observed groups, or by shuffling edges among individuals observed in the same location at the same time, one can generate a large number of permuted networks that have the same structural biases as the collected data but lack any biological processes that would cause additional non-random patterns. The difference in the number of connections (degree) of males and females, for example, could be compared between the observed and randomised networks, indicating whether males are interacting with more other individuals than females, given their distribution among groups. These permutations can be further constrained to account for patterns of interactions that might arise from heterogeneously distributed resources (Ramos-Fernández et al. 2006), or other factors not related specifically to the social tendencies of individuals. Such permutations are very common, and well described in primers and "How-to" guides (Farine 2013, 2017; Farine and Whitehead 2015).

An alternative approach is to fit statistical models developed for use in networks directly to the observed network data. Examples of these models include exponential random graph models (ERGMs; Lusher, Koskinen, & Robins, 2012; Robins, Pattison, Kalish, & Lusher, 2007) and stochastic actor-oriented models (Snijders et al. 2010; Ilany et al. 2015; Fisher et al. 2017a), which have both

been applied previously to analyse animal networks. With these approaches, terms, similar to those fitted in a linear model, are specified to model the probability or weight of edges in the networks. These terms can explicitly relate to other links in the network, hence directly modelling the non-independence of network data. An additional benefit is that the nature of the dependence assumption made can be specified within the model (Robins et al. 2007; Lusher et al. 2012), although this does add complexity to model implementation. Further terms can be fitted that represent factors that may underly differences in social behaviour, for example, for individuals of a certain type (e.g. individuals of the same sex) to associate more or less (Silk and Fisher 2017). Using ERGM parameters for explanatory variables in the model are estimated simultaneously, for example estimating the difference in the number of connections between males and females, while accounting for the fact that individuals may be in different groups or live varying distances apart. Simultaneous estimation allows one to evaluate multiple competing hypotheses for the formation of animal social structure, while controlling for potentially confounding factors (Desmarais and Cranmer 2012). In addition, because ERGMs are fitted to the observed network itself, they provide a more direct measure of the importance of combinations of covariates in explaining social structure. However, some authors have suggested that ERGM parameter estimates may be sensitive to missing data (Shalizi and Rinaldo 2013), and their performance when analysing data collected through group-membership has not yet been thoroughly tested (Farine 2017; Silk and Fisher 2017).

Permutation and ERGM approaches are distinct approaches, yet often can be used in the same way to test hypotheses about the structure of animal social networks. Despite this, they have not been simultaneously evaluated in the context of analysing animal social network data. This means that there is a paucity of information on how relatively well each approach performs for different types of network, methods of data collection, or questions in animal SNA. On one hand, generative network models such as ERGMs have been designed for studies of human social networks. This means that ERGMs may not be appropriate to model some animal social network data, as such networks are often based on inferred relationships, missing data can be a considerable problem, and there may be great biases generated by the method of data collection. On the other hand, permutation-based approaches require appropriate, and often system-specific, null models and their performance might depend on other features of the network in ways that are challenging to predict.

We assessed the performance of both permutation-based and ERGM approaches to test hypotheses relating individual traits to the strength of social network connections in simulated network data. The relationship between individual traits and network connectivity is a common research question in studies of animal social networks for which both of these approaches are

appropriate. For some network-related hypotheses (e.g. the consistency of individual position within networks for different behaviours or time periods, or when the network trait is a predictor variable), ERGMs are less applicable and other approaches should be used. We used simulated data, rather than real data, for two key reasons: a) we could control the "biological" signal in the datasets, and so we knew the true effect and could assess whether either method accurately recovered it (e.g. Bonnet and Postma 2016); and b) we had a close underlying understanding of the generative processes underlying our emergent network structures, meaning that we could more effectively explain variation in model performance.

We simulated networks that varied considerably in their structure and sampling methodology to recreate a diversity of network types likely to be encountered in animal network analysis. We simulated two broad types of network: dyadic-based (for interaction or contact networks) and grouping-event based (sometimes termed association) networks. Our aim was not to compare these different kinds of network, but to simultaneously evaluate the performance of both ERGMs and permutation-based approaches when analysing them. Our dyadic-based networks represent the types of networks constructed by researchers using data from proximity loggers or direct observations of behavioural interactions between individuals. Such data might be gathered by researchers collecting data on terrestrial mammals using proximity loggers, or aggressive interactions between individually marked fiddler crabs. Our grouping-event based networks represent the types of networks constructed by researchers using the Gambit of the Group assumption (Whitehead and Dufault 1999), where individuals overlapping in space and time are deemed to have associated. Such data might be gathered by researchers observing flocks of ringed birds or shoals of tagged fish. We also manipulated other parameters in our network generation process, enabling us to vary other key aspects of animal network structure such as modular structure (common in group-living or fission-fusion societies) and the importance of space in determining connectivity in the network.

Once we had simulated these dyadic- and grouping event-based data, we then sampled them with a range of sampling intensities, to give us data sets analogous to those collected by animal social network researchers. We looked for a sex effect on an individual's network "strength", which is the sum of all an individual's weighted edges in the network. Using strength as a response variable represents a researcher testing a biologically plausible hypothesis (e.g. females have more and/or stronger connections than males). We used a range of parameter values that resulted in either no difference between the sexes, more gregarious males, or more gregarious females. We also added various confounding effects to our networks, for instance the presence of positive or negative assortativity by sex, or stronger or weaker effects of distance between individuals. We then

analysed the networks with each approach and measured the frequency of false positive (type I) and false negative (type II) errors. We predicted that permutation-based approaches would outperform ERGM-based approaches in networks with a high density of edges, in particular in grouping event based networks with sampling error (Farine 2017). However, we anticipated that ERGM approaches would perform better in dyadic networks, especially those with lower edge densities, as a result of directly incorporating confounding effects.

Methods

Our methods comprised of three stages: initial network generation to generate the underlying social structure of the population, network sampling to generate the two different types of social network data, and network analysis.

1. Network generation

We simulated social networks to emulate patterns of interactions seen in real networks. The frequency of interactions depended on the sexes of both members of the dyad. Males could be generally more, equally or less social than females (Wolf et al. 2007). Detecting this effect was our test of the models' performance. Frequency of interactions between individuals could also depend on whether individuals were the same sex, part of the same social group and on the distance between their groups. Intra-sex interactions could therefore be more, the same, or less strong than inter-sex interactions. Similarly, within-group interactions could be as or more common than among-group interactions (Weber et al. 2013), and interactions between closer individuals could be as or more common than those further apart (Best et al. 2014). These non-random elements of our simulations create confounding signal within the networks which may influence the analysis.

Detailed methods:

For each network we generated a population of 100 individuals of random sex, randomly sorted into 10 groups of 10. Each group was assigned a random location in space. Distance between these locations was normalised so that the greatest possible distance was 1. Dyadic associations were potentially generated between all individuals in the population based on their sex, whether the interaction was within or between groups and the distance between the groups. Specifically, for each dyad, edge weight was the sum of two integers, each drawn randomly from the following negative binomial distribution:

$$NB\left(size = (m.i.eff + g.dens) \times dist^{d.eff}, prob = 0.3\right)$$

Where *m.i.eff* is the effect of being male (always 0 for females), *g.dens* is controls the baseline strength of interactions, the value of which is dependent on whether an interaction is within a group (*i.dens* in Table 1) or between groups (*o.dens* in Table 1). *dist* is the inverted distance between groups (so that 1 is within the same group and 0.001 is the greatest distance between groups) and *d.eff* is modifier for the effect of distance. Each individual in a dyad therefore has a value generated from their own negative binomial distribution (see supplementary Figs. S1 & S2). These values are then summed to obtain the weight of the edge connecting that dyad. The weights of edges between the same sex were then multiplied by an additional term, *sex.eff* to increase or decrease the frequency of same sex interactions. For each combination of these parameters (a total of 243, see Table 1) we generated 100 undirected, weighted adjacency matrices. We refer to these as the "true" networks (Figure 1).

2. Network sampling

Having generated the true network, we then simulated two different methods by which researchers might attempt to measure these relationships. First, we simulated dyad-based networks, as might be generated by observations of behavioural interactions (e.g. grooming), or bio-logging data (e.g. proximity loggers). Secondly, we simulated grouping event-based networks, in which all individuals observed associating in a single grouping event are assumed to have engaged in a biologically meaningful social interaction (Whitehead and Dufault 1999).

Social network data collected on animals are often far from complete: unidentified individuals often make up considerable portions of populations, and many interactions and grouping events simply go unobserved (Franks et al. 2009, 2010; Farine 2014; Silk et al. 2015; Davis et al. 2018). We therefore simulated our measurements at differing accuracies, governed by an observation effort parameter. The observation effect parameter had the values of 0.3, 0.6, 0.9 and 1, where 1 is considered complete sampling of either the network or the series of grouping events used to construct it (See Fig. 1 for a network diagram showing the effect of differing observation effects and network types). Introducing sampling effects may create opportunities for spurious effects to be detected (e.g. incomplete data may create the impression that individuals prefer to associate with individuals with a same number of connections as them, when no such effect exists in the network), but also prevent real effects from being detected.

Detailed methods:

Dyad-based networks:

Dyad-based networks were generated by adding noise to the true network. For each edge, a new edge weight was randomly selected from a sequence ranging from zero to the "true" edge weight, using a probability distribution where the higher the observation effort, the greater the probability that the value selected would be closer to the true edge weight. Edge weights (for permutation based and ERGM approaches) therefore remained un-scaled counts of interactions as would be expected from networks of dyadic interactions or counts of contacts based on proximity. The simulated error may represent hardware problems or missed observations. For graphical illustration of how observation effort affects the likelihood of choosing the true edge weight, see supplementary Fig. S3.

Grouping event-based networks:

For each true network we generated a group-by-individual matrix (GBI: recording which individuals were recorded in a given grouping event) consisting of 1000 grouping events (n.b. grouping events are distinct from the group membership of individuals in the underlying network of true social relationships). To generate a grouping event, a random individual was chosen from the population to act as the "seed" of the grouping event (Fig. S4a). Edge weights were rescaled between 0 and 1 – where 1 was the greatest edge weight in the true network. The squared, rescaled dyadic edge weights of the "seed" individual with all other members of the population were used as the probability of success in a random binomial trial. Any individuals with successes were added to the grouping event (Fig. S4b). As we defined a grouping event as consisting of at least two members, this process was repeated until at least one other individual was added to the event.

After generating a grouping event, each member of the event was then used (one at a time, in a random order) as focal individual (Fig. S4c). Further members were added to the event based on the strength of their connections with the focal individual. Unlike when generating the event, here it was possible for no individuals to be added to the event when considering a focal individual. At this stage, if a potential joiner had an edge of weight zero with any individual already in the event, the probability of the potential joiner being added to the event was reduced to 0.01, regardless of the strength of the connection to the current focal individual (Fig. S4c and d). This represents the potential individual being unlikely to be part of this grouping event due to the presence of members with whom they have no connection in the true network, but with a small chance that these individuals could occur within the same group. Each group member added to an event was treated as a focal individual themselves until every member had been treated as a focal individual (Fig. S4d).

Once all 1000 grouping events had been generated, a proportion of these events were randomly discarded depending on observation effort (the proportion equalling 1 – $obs.eff$). These represented unobserved grouping events. The remaining GBI matrices were then converted into adjacency matrices, with edge weight being the number of grouping events two individuals co-occurred in. For the permutation-based analysis of the grouping event-based networks edge weights consisted of the simple ratio index (Cairns and Schwager 1987) - the number of grouping events in which a pair of individuals were observed together was divided by the sum of the number of events each individual was observed in. For the ERGM-based analysis edge weights consisted of the number of groups individuals were seen in together, to be consistent with the type of ERGM we fitted.

Networks generated using grouping event-based approaches can create subtly differently structured networks (Franks et al. 2010). We confirmed that both the group-based and dyad-based networks generated using our algorithm were broadly representative of the true network using Mantel tests (Mantel 1967) during the development of these simulations (see Fig. S5 for results of these Mantel tests, Fig. 1 for a network diagram comparing the true network with the sampled network and Figs. S6 – 8 for similar figures for further parameter sets).

3. Network analyses

We assume for the purposes of this analysis that the researcher approaching these network data is not specifically interested in how individuals are assorted within vs. among groups, or within vs. between the sexes, but that they acknowledge that this occurs in their study system. Instead, they wish to determine whether males and females differ in their frequency and strength of their social relationships.

ERGMs treat the network as a response variable and fit parameter by finding values that produce sets of edges with similar properties to those in the observed network (Robins et al. 2007; Hunter et al. 2008). Initially they were developed to model the presence/absence of edges as binary response variables, but subsequent developments have facilitated the development of ERGMs for weighted networks (Lusher et al. 2012). For our ERGMS, we fitted a count ERGM to the networks (Krivitsky 2012, 2015), as our association strengths are integers. For the dyadic networks, we fitted a term for “sex assortativity”, modelling the tendency for individuals of the same sex to interact more or less frequently, and “same-group”, modelling the tendency for individuals within the same group to interact more frequently. We also fit the distance between each dyad (based on the location of their groups), an *n* x *n* matrix, as a dyadic covariate, modelling the tendency for individuals living further apart to interact less. For the grouping event-based networks, as the data were collected by observing many grouping events and “true” group membership was assumed to be unknown, we did

not fit a term for shared group membership but did include a dyadic covariate that consisted of a distance matrix for home range centroids. Each individual's home range centroid was calculated as the mean location of the groups the individual was observed in. To detect the biological signal of interest in both types of network, we included a term for sex-degree to investigate the tendency for the sexes to have a different level of gregariousness. We confirmed a subset of models had converged and fitted the networks appropriately following Lusher et al. (2012). We considered the model to have detected an effect when $p < 0.05$.

For the permutation-based approach, we generated permuted networks in one of two ways. For each dyad-based network, we simulated 10,000 networks where the rows and columns of the dyad-based network were shuffled using the "rmperm" function in the R package sna (Butts 2008). For the grouping-event based networks, we created 10,000 permutations of each network using the function "network_swap" in the package asnipe (Farine 2013). This permutes the data stream by swapping individuals between grouping events 10,000 times, resulting in 10,000 randomised networks. We constrained these swaps to only occur between individuals within the same location, to account for of the effect of space on network structure. We then constructed a new network for each permutation.

In each of our dyad- and grouping event-based networks and the permuted versions of these, we compared the weighted degree of males and females using a (G)LMs. We used a Poisson error distribution for dyad-based networks and a Gaussian error distribution for grouping event-based networks due to the differences in edge weights between the two (edge weights of dyad-based networks were counts and edge weights of grouping-event based networks used the simple ratio index for the permutation-based analysis). We compared the distribution of effect sizes from the permuted networks to the effect size from the observed network (Farine 2017). P-values were calculated as the proportion of effect sizes in the permuted networks that were smaller than the effect size in the observed network. We considered the model to have detected an effect when $p < 0.05$ (in a two-tailed test). These comparisons allowed us to determine whether the differences in weighted degree between the sexes, differed from that expected in the permuted networks To calculate the rate of false positives, for the 100 networks in each parameter set, where the effect of being male was set at 0, we counted the number of times the model detected a difference between the sexes in weighted degree. This gives a failure rate out of 100. To calculate the rate of false negatives, for the 100 networks in each parameter set where the effect of being male was not 0, we counted the number of times the model failed to detect a difference between the sexes in weighted degree. This also gives a failure rate out of 100. We examined how the rates of false positives and negatives vary depending on each level of our other parameters (the *sex effect*, *within-group edge*

density, *between-group edge density*, *distance effect* and *observation effort*). For all parameters other than the observation effort, we only consider cases when the *observation effort* was 1.

Results

We provide an overview of key findings in the main text based on graphs of the error rates of the two methods under different scenarios. For the number and percentage of simulations with error rates over 5% and 10% for each of the levels of the parameters plotted here, please see Tables S2-S5 in the supplementary materials.

False positives

Both ERGMs and our permutation-based approach were relatively prone to false positives in dyad-based and grouping event-based networks (Fig. 2, columns a and b). False positive rates (at $\alpha = 0.05$) were typically lower for ERGMs than for permutations, and lower in dyadic networks than in grouping event-based networks.

Dyadic networks

The difference in false positives was marginal for dyadic networks, with false positive rates typically lower for ERGMs than for permutations (Fig. 2a). The presence of a confounding effect of assortativity by sex had the greatest effect on rates of false positives compared with other parameters tested. The permutation-based approached performed relatively well when there was no assortment by sex but poorly otherwise. In contrast, ERGMs performed best when the network was negatively assorted by sex, and worst when positively assorted by sex (Fig. 2a i). While the performance of ERGMs was unaffected by any other parameters, including the density of within group interactions (Fig. 2a ii) the permutation-based approach performed worse when there was a higher density of between-group connections (Fig. 2a iii) or with a distance effect of zero (Fig. 2a iv), i.e. in situations when the group structure of the network was less clear.

Grouping event-based networks

Both ERGMs and permutation-based methods produced a high false positive rate of around 40% in grouping-event-based networks (Fig. 2b). ERGMs showed a much more variable error rate than the permutation-based approach, which was quite consistent. Similar to the results for dyad-based networks, ERGMs performed best with negative assortativity by sex and worst with positive

assortativity, while permutations performed best with no assortativity by sex (Fig. 2b i). However, unlike the results for dyad-based networks, permutations also performed well when there was negative assortativity, while ERGMs performed nearly as poorly under no assortativity as under positive assortativity. Changes to network structure had different impacts on false positive rates for ERGMs and permutations. Increasing both the within- and between-group edge density increased the false positive rate for ERGMs (Figs. 2b ii and 2b iii). For permutations there was a smaller effect, with a slight reduction in false positive rates when within-group density increased (Figs. 2b ii and 2b iii). Increasing the distance effect had relatively little effect on the rates of false positives for both ERGMs and permutations (Fig. 2b iv).

False negatives

Both ERGMs and our permutation-based approach were much less prone to false negatives than false positives. The rates of false negatives were especially low in dyad-based networks and higher in grouping event-based networks (Fig. 2, columns c and d). False negative rates were typically lower for ERGMs than for permutations.

Dyad-based networks

Both methods were highly effective at detecting differences in weighted degree between the sexes and had very low rates of false negatives in dyad-based networks (Fig. 2c). This was generally true whether the assortativity effect was positive, negative or absent, although both methods, especially the permutation-based approach, showed an increase in false negative rates when the networks were negatively assorted by sex (Fig. 2c i). Both methods had higher false negative rates when the within-group edge density was lower (Fig. 2c ii), but between-group edge density had no clear effect (Fig. 2c iii). When distance had a stronger negative effect on between group edges (i.e. connections among members of distant groups were highly unlikely) both methods had slightly reduced performance (Fig. 2c iv).

Grouping event-based networks

Both methods produced false negative rates of approximately 10% for ERGMs and 20% for permutations, higher than for dyad-based networks (Fig. 2d). False negative rates for ERGMs and permutations were much higher when networks were negatively assorted by sex than when they were not assorted or positively assorted (Fig. 2d i). Networks with stronger within-group connections

had lower rates of false negatives for both methods (Fig. 2d ii), with stronger between-group connections having a similar but smaller effect on the false negative rate of ERGMs only (Fig. 2d iii). As for dyad-based networks, increasing the strength of the distance effect on between-group connections increased false negative rates for both methods (Fig. 2d iv).

The effect of network sampling on error rates

Sampling a subset of possible interactions or contacts from the dyad-based network in an unbiased manner had no clear effect on rates of either false positives (Fig. 3a) or false negatives (Fig. 3b). In contrast, sampling a subset of possible grouping events had a considerable effect on inference in grouping event-based networks. Contrary to our predictions, there were more false positives when a more complete sample of grouping events conducted (Fig. 3c) while, conversely, increased observation effort reduced the rate of false negatives (Fig. 3d).

Discussion

We have evaluated the performance of both ERGM and permutation-based approaches for analysing animal social networks in a range of contexts. There are four key take-home messages from our work. First, ERGMs generally performed well, producing low rates of false positives for dyad-based networks, and lower rates of false negatives in both dyad- and grouping event-based networks. Second, both ERGMs and datastream permutations had high false positive rates in grouping-event based networks, supporting similar results from Weiss et al. (2020) for permutations and highlighting that ERGMs do not necessarily provide a viable alternative in this context without careful consideration of additional variables to control for sampling effects. Third, the performance of both approaches depended on the assortativity of the network; both approaches performed well when there was no assortativity by sex, permutation-based approaches performed poorly when there was any assortativity by sex and ERGMs performed poorly when there was positive assortment by sex. Fourth, in grouping event-based networks both analysis approaches gave lower rates of false negatives, but higher rates of false positives, as observation effort increased. These results should aid researchers in choosing appropriate analytical approaches in animal social network studies. We have summarised our key findings and recommendations in Table 2. We stress however that no network analysis method is “plug and play”, and that careful consideration should be given when fitting an ERGM or when designing permutations to analyse any network.

In dyad-based networks, rates of false positives were relatively low for detecting differences in degree, although typically above those that would be expected for $\alpha = 0.05$. False positive rates were typically higher for permutations than ERGMs for all dyadic networks. Permutations may therefore be anti-conservative when analysing dyad-based networks. This was particularly true when additional effects are present in networks, as ERGMs performed better than permutations if there was either positive or negative assortativity by sex, but not if there was no sex-assortativity. Performance was worst for both methods when connections were positively assorted by sex, while permutations also performed badly when networks were negatively assorted by sex, yet ERGMs performed best in this context. The poor performance of permutations in this context suggests that when a trait affects both degree and assortativity, permutation-based approaches are more likely to detect spurious differences between categories of individuals (such as male and female) in their number of connections. This highlights the benefits of using ERGMs over permutation-based approaches in this context; namely that ERGMs can more easily facilitate the incorporation of additional confounding variables when testing an effect of interest as ERGMs specifically model topological effects on network structure alongside other biological processes of interest (Silk and Fisher 2017). A caveat here is that there were differences in ERGM error rates that depended on whether assortativity was positive or negative. This reveals that assortativity may influence network structure in a way that alters model performance even when accounted for. Phenotypic assortativity is common in animal social networks across taxa and for a range of different traits (Farine 2014; McDonald and Pizzari 2016; McDonald et al. 2017). We therefore suggest that caution is applied when testing for differences in connectivity or social centrality in study systems in which such patterns of assortativity are expected to occur. Positive assortativity (e.g. males interacting more with other males) will often cause a difference in connectivity to be found when it is in fact absent, while negative assortativity (e.g. males being more likely to interact with females) can lead to a difference in connectivity being missed when they are present. Future work to develop approaches that can better address these biases in estimation will be valuable.

In grouping event-based networks the two analysis approaches did not greatly differ in overall effectiveness but did show different patterns. Stronger within- and between-group interactions increased false positive rates for ERGMs but decreased them for the permutation approach. In other words, more network connections increased the chance that ERGMs would detect an effect when there was none, but fewer network connections increased the chances the permutation approach would correctly identity no effect. In contrast, increasing the density of within-group interactions or reducing the distance effect so that networks were more widely connected decreased the false negative rate in grouping event-based networks for both approaches.

A difference between the sexes was therefore easier to detect in grouping event-based networks that were more well-connected. While for ERGMs this represents a trade-off between false positives and false negatives as the number of complete edges increases, permutations will perform consistently better in well-connected networks compared to sparsely connected networks , with relatively lower rates of both false positives and false negatives.

Lower levels of observation effort increased the rate of false negatives in grouping event-based networks, with this effect especially striking when only 30% of groups were sampled (when 60% of groups were sampled, error rates were more similar to full sampling). This highlights that under-sampling grouping events may lead to inaccurate inferences as reported elsewhere (Franks et al. 2010; Farine 2014; Fisher et al. 2017b), especially when many grouping events are missed (and the number sampled is low). Interestingly, increasing the observation effort *increased* the rate of false positives in grouping-event based networks. Therefore, for both approaches a higher number of observed interactions (dyad-based networks) or grouping events (grouping event-based networks) increases the chances of an effect being found, regardless of whether it was actually present. A similar effect was found for datastream permutations by Weiss et al. (2020), with false positives increasing as more grouping events were sampled. We suggest that for permutation-based approaches, the problems associated with datastream permutations highlighted by Weiss et al. (Weiss et al. 2020) are exacerbated when observation effort is higher. When more events are sampled, the randomisation process results in permuted networks that have less variation in connectivity and edge weight than when fewer grouping events are sampled. Why this also happens for ERGMs as well is less clear, although does support the suggestion of Shalizi and Rinaldo (Shalizi and Rinaldo 2013) that in some contexts ERGMs may be susceptible to sampling effects. As a result, those studying dense grouping-event based social networks should be cautious when interpreting any statistically significant effects they detect, as the effects could be spurious. These effects were absent in dyad-based networks, suggesting that they are tied to type of sampling used. Future work could explore the impact of sampling in more detail to produce a sensitivity curve for the effect of sampling effort on error rates in animal social network studies that exploit data on group membership.

Moving forward, edge weights that represent residuals of models that account for space use (Whitehead and James 2015) might represent useful approaches to study population-level social networks. A further alternative may be to use network models that can control for space more effectively such as latent space models (Cranmer et al. 2016; Silk et al. 2017). Latent space models deal with the non-independence of individuals in a network by placing them within a k-dimensional "social space" (Hoff et al. 2002), and this is likely to handle individuals with different sets of contacts

more effectively than either approach used here. Further developments of permutation and ERGM approaches will also be possible. The use of bipartite ERGMs to directly model group by individual matrices offers one potential solution for grouping event-based networks (Silk et al. 2017). However, it may also be possible to fit additional terms in count-based ERGMs, or use alternative edge weight distributions, to control for sampling effects. Similarly, the use of datastream permutations that can maintain key network features (such as degree distributions), similar to those suggested by Chodrow (2019), might reduce the false positive rates of these approaches in grouping event-based networks. Using such datastream permutations may be especially beneficial if these approaches are combined with more conventional biological constraints (Whitehead and Dufault 1999; Whitehead et al. 2005; Croft et al. 2011; Farine and Whitehead 2015; Farine 2017). However, it must be confirmed that these approaches do not suffer the same problems as those identified by Weiss et al. (Weiss et al. 2020).

Conclusions

In conclusion, we have examined the relative strengths and weaknesses of applying ERGMs and permutation-based approaches in a range of animal social networks in the presence and absence of confounding effects. Our study, alongside other works investigating how best to statistically examine and interpret animal networks, provide a series of guidelines for empiricists moving forward (Table 2). Overall, while both ERGM and permutation-based approaches have their weaknesses, both clearly offer valuable tools in analysing animal social networks, and further methodological developments that improve the performance of both in grouping event-based data should be a priority.

Acknowledgements

JCE thanks Julie Morand-Ferron, DNF Andrew McAdam and Jonathan Pruitt, and MJS Robbie McDonald and Dave Hodgson for not minding them pursuing this project in lieu of their post-doc research. There is no specific funding to report. We have no conflicts of interest.

Data Accessibility

The R code used to simulate and analyse the networks are available as supplemental files. Simulation R code, and necessary summary data and R code to reproduce the analyses reported in this article are provided by (Evans et al. 2020).

Authors' contributions

All authors came up with the initial idea for the paper. JCE and MJS wrote the code to generate and analyse the networks with input from DNF. All authors contributed to writing the manuscript and approved of the final draft.

References

Anderson, B. S., C. Butts, and K. Carley. 1999. The interaction of size and density with graph-level indices. Soc. Networks 21:239–267.

Bautista, A., J. A. Zepeda, V. Reyes-Meza, M. Martínez-Gómez, H. G. Rödel, and R. Hudson. 2015. Contribution of within-litter interactions to individual differences in early postnatal growth in the domestic rabbit. Anim. Behav. 108:145–153.

Bejder, L., D. Fletcher, and S. Bräger. 1998. A method for testing association patterns of social animals. Anim. Behav. 56:719–725.

Berman, and Kapsalis. 1999. Development of kin bias among rhesus monkeys: maternal transmission or individual learning? Anim. Behav. 58:883–894.

Best, E. C., R. G. Dwyer, J. M. Seddon, and A. W. Goldizen. 2014. Associations are more strongly correlated with space use than kinship in female eastern grey kangaroos. Anim. Behav. 89:1–10. Elsevier.

Bonnet, T., and E. Postma. 2016. Successful by Chance? The Power of Mixed Models and Neutral Simulations for the Detection of Individual Fixed Heterogeneity in Fitness Components. Am. Nat. 187:60–74. University of Chicago PressChicago, IL.

Butts, C. C. T. 2008. Social network analysis with sna. J. Stat. Softw. 24:13–41.

Cairns, S. J., and S. J. Schwager. 1987. A comparison of association indices. Anim. Behav. 35:1454–1469.

Cheney, D. L., J. B. Silk, and R. M. Seyfarth. 2016. Network connections, dyadic bonds and fitness in wild female baboons. R. Soc. Open Sci. 3:160255.

Chodrow, P. S. 2019. Configuration Models of Random Hypergraphs and their Applications. arXiv Prepr. arXiv1902.09302.

574 Clutton-Brock, T. H., K. E. Rose, and F. E. Guinness. 1997. Density-related changes in sexual selection
575 in red deer. Proc. Biol. Sci. 264:1509–1516.

576 Cranmer, S. J., P. Leifeld, S. D. McClurg, and M. Rolfe. 2016. Navigating the range of statistical tools
577 for inferential network analysis. Am. J. Pol. Sci. Wiley Online Library.

578 Croft, D. P., R. James, and J. Krause. 2008. Exploring animal social networks. Princeton University
579 Press.

580 Croft, D. P., J. Krause, and R. James. 2004. Social networks in the guppy (Poecilia reticulata). Proc.
581 Biol. Sci. 271 Suppl:S516-9.

582 Croft, D. P., J. R. Madden, D. W. Franks, and R. James. 2011. Hypothesis testing in animal social
583 networks. Trends Ecol. Evol. 26:502–507. Elsevier.

584 Davis, G. H., M. C. Crofoot, and D. R. Farine. 2018. Estimating the robustness and uncertainty of
585 animal social networks using different observational methods. Anim. Behav. 141:29–44.
586 Academic Press.

587 Desmarais, B. A., and S. J. Cranmer. 2012. Statistical inference for valued-edge networks: the
588 generalized exponential random graph model. PLoS One 7:e30136. Public Library of Science.

589 Evans, J., D. N. Fisher, and M. J. Silk. 2020. Silk, Matthew; Evans, Julian; Fisher, David The
590 performance of permutations and exponential random graph models when analysing animal
591 networks (R code and data). https://doi.org/10.5061/dryad.9w0vt4bcn.

592 Farine, D. R. 2017. A guide to null models for animal social network analysis. Methods Ecol. Evol.
593 8:1309–1320. Wiley Online Library.

594 Farine, D. R. 2013. Animal social network inference and permutations for ecologists in R using
595 asnipe. Methods Ecol. Evol. 4:1187–1194. Wiley Online Library.

596 Farine, D. R. 2014. Measuring phenotypic assortment in animal social networks: weighted
597 associations are more robust than binary edges. Anim. Behav. 89:141–153. Elsevier.

598 Farine, D. R., and H. Whitehead. 2015. Constructing, conducting and interpreting animal social
599 network analysis. J. Anim. Ecol. 84:1144–1163. Wiley Online Library.

600 Fewell, J. H. 2003. Social insect networks. Science 301:1867–70.

601 Fisher, D. N., A. Ilany, M. J. Silk, and T. Tregenza. 2017a. Analysing animal social network dynamics:
602 the potential of stochastic actor-oriented models. J. Anim. Ecol. 86:202–212.

603 Fisher, D. N., M. J. Silk, and D. W. Franks. 2017b. The perceived assortativity of social networks:
604 Methodological problems and solutions. Pp. 1–19 *in* Lecture Notes in Social Networks.

605 Frank, S. A. 2007. All of life is social. Curr. Biol. 17:R648–R650.

606 Franks, D. W., R. James, J. Noble, and G. D. Ruxton. 2009. A foundation for developing a
607 methodology for social network sampling. Behav. Ecol. Sociobiol. 63:1079–1088.

608 Franks, D. W., G. D. Ruxton, and R. James. 2010. Sampling animal association networks with the
609 gambit of the group. Behav. Ecol. Sociobiol. 64:493–503. Springer.

610 Hoff, P. D., A. E. Raftery, and M. S. Handcock. 2002. Latent Space Approaches to Social Network
611 Analysis. J. Am. Stat. Assoc. 97:1090–1098. Taylor & Francis.

612 Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. 2008. ergm: A package to
613 fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw.
614 24:nihpa54860. NIH Public Access.

615 Ilany, A., A. S. Booms, and K. E. Holekamp. 2015. Topological effects of network structure on long-
616 term social network dynamics in a wild mammal. Ecol. Lett. 18:687–695. Wiley Online Library.

617 James, R., D. P. Croft, and J. Krause. 2009. Potential banana skins in animal social network analysis.
618 Behav. Ecol. Sociobiol. 63:989–997.

619 Krause, J., R. James, D. W. Franks, and D. P. Croft. 2014. Animal social networks. Oxford University
620 Press.

621 Krivitsky, P. N. 2015. ergm.count: Fit, Simulate and Diagnose Exponential-Family Models for
622 Networks with Count Edges.

623 Krivitsky, P. N. 2012. Exponential-family random graph models for valued networks. Electron. J. Stat.
624 6:1100. NIH Public Access.

625 Leu, S. T., J. Bashford, P. M. Kappeler, and C. M. Bull. 2010. Association networks reveal social
626 organization in the sleepy lizard. Anim. Behav. 79:217–225.

627 Lusher, D., J. Koskinen, and G. Robins. 2012. Exponential Random Graph Models for Social Networks:
628 Theory, Methods, and Applications. Cambridge University Press.

629 Lusseau, D. 2003. The emergent properties of a dolphin social network. Proc. R. Soc. London. Ser. B
630 Biol. Sci. 270:S186–S188. The Royal Society.

631 Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer

632 Res. 27:209–220.

633 McDonald, G. C., D. R. Farine, K. R. Foster, and J. M. Biernaskie. 2017. Assortment and the analysis of
634 natural selection on social traits. Evolution (N. Y). 71:2693–2702. Wiley/Blackwell (10.1111).

635 McDonald, G. C., and T. Pizzari. 2016. Why patterns of assortative mating are key to study sexual
636 selection and how to measure them. Behav. Ecol. Sociobiol. 70:209–220. Springer.

637 Myers, J. P. 1983. Space, time and the pattern of individual associations in a group-living species:
638 Sanderlings have no friends. Behav. Ecol. Sociobiol. 12:129–134.

639 Ramos-Fernández, G., D. Boyer, and V. P. Gómez. 2006. A complex social structure with fission–
640 fusion properties can emerge from a simple foraging model. Behav. Ecol. Sociobiol. 60:536–
641 549.

642 Robins, G., P. Pattison, Y. Kalish, and D. Lusher. 2007. An introduction to exponential random graph
643 (p*) models for social networks. Soc. Networks 29:173–191. Elsevier.

644 Sade, D. S. 1972. Sociometrics of Macaca mulatta. I. Linkages and cliques in grooming matrices. Folia
645 Primatol. (Basel). 18:196–223.

646 Shalizi, C. R., and A. Rinaldo. 2013. Consistency under sampling of exponential random graph
647 models. Ann. Stat. 41:508. NIH Public Access.

648 Silk, M. J., D. P. Croft, R. J. Delahay, D. J. Hodgson, N. Weber, M. Boots, and R. A. McDonald. 2017.
649 The application of statistical network models in disease research. Methods Ecol. Evol., doi:
650 10.1111/2041-210X.12770.

651 Silk, M. J., and D. N. Fisher. 2017. Understanding animal social structure: exponential random graph
652 models in animal behaviour research. Anim. Behav. 132.

653 Silk, M. J., A. L. Jackson, D. P. Croft, K. Colhoun, and S. Bearhop. 2015. The consequences of
654 unidentifiable individuals for the analysis of an animal social network. Anim. Behav. 104:1–11.
655 Elsevier Ltd.

656 Snijders, T. A. B., G. G. Van de Bunt, and C. E. G. Steglich. 2010. Introduction to stochastic actor-
657 based models for network dynamics. Soc. Networks 32:44–60. Elsevier.

658 Strandburg-Peshkin, A., D. R. Farine, M. C. Crofoot, and I. D. Couzin. 2017. Habitat and social factors
659 shape individual decisions and emergent group structure during baboon collective movement.
660 Elife 6:e19505. eLife Sciences Publications Limited.

661 Sumpter, D. J. T. 2006. The principles of collective animal behaviour. Philos. Trans. R. Soc. Lond. B.
662 Biol. Sci. 361:5–22.

663 Wasserman, S., and K. Faust. 1994. Social Network Analysis: Methods and Applications. Cambridge
664 University Press, Cambridge.

665 Webber, Q. M. R., and E. Vander Wal. 2019. Trends and perspectives on the use of animal social
666 network analysis in behavioural ecology: a bibliometric approach. Anim. Behav. 149:77–87.
667 Elsevier.

668 Weber, N., S. P. Carter, S. R. X. Dall, R. J. Delahay, J. L. McDonald, S. Bearhop, and R. A. McDonald.
669 2013. Badger social networks correlate with tuberculosis infection. Curr. Biol. 23:R915–R916.
670 Elsevier.

671 Weiss, M. N., D. W. Franks, L. J. N. Brent, S. Ellis, M. J. Silk, and D. P. Croft. 2020. Common
672 datastream permutations of animal social network data are not appropriate for hypothesis
673 testing using regression models. bioRxiv 2020.04.29.068056. Cold Spring Harbor Laboratory.

674 Whitehead, H., L. Bejder, and C. Andrea Ottensmeyer. 2005. Testing association patterns: issues
675 arising and extensions. Anim. Behav. 69:e1.

676 Whitehead, H., and S. Dufault. 1999. Techniques for analyzing vertebrate social structure using
677 identified individuals: review and recommendations. Adv. Study Behav. 28.

678 Whitehead, H., and R. James. 2015. Generalized affiliation indices extract affiliations from social
679 network data. Methods Ecol. Evol. 6:836–844. Wiley Online Library.

680 Wittemyer, G., I. Douglas-Hamilton, and W. M. Getz. 2005. The socioecology of elephants: analysis of
681 the processes creating multitiered social structures. Anim. Behav. 69:1357–1371.

682 Wolf, J. B. W., D. Mawdsley, F. Trillmich, and R. James. 2007. Social structure in a colonial mammal:
683 unravelling hidden structural layers and their foundations by network analysis. Anim. Behav.
684 74:1293–1302. Elsevier.

685

686 ## Tables

687

688 Table 1. Parameters of interest and the values used in network generation and sampling.

Name	Description	Values	Values description
d.eff	Effect of distance between groups on the frequency of between groups-interaction	0	Distance between groups has no effect
		4	Increased distance reduces likelihood of interaction moderately
		8	Increased distance reduces likelihood of interaction strongly
i.dens	Effect of an interaction being within a group	0.4	Interactions within groups less common
		0.8	Interactions within groups quite frequent
		1.2	Interactions within groups very frequent
o.dens	Effect of an interaction being between groups	0.4	Interactions between groups less common
		0.2	Interactions between groups rare
		0.1	Interactions between groups extremely rare
m.eff	Effect of being male	-0.5	Males less likely to be involved in social interactions
		0	Being male has no effect on frequency of interactions
		0.5	Males more likely to be involved in social interactions
sex.eff	Strength of intra-sex interactions	0.5	Intra sex interactions weaker
		1	No effect of intra-sex interactions
		2	Intra sex interactions stronger
obs.eff	Observation effort	0.3	Lazy observer
		0.6	Diligent observer
		0.9	Superhero observer
		1.0	Omniscient observer

689

690

691

692 Table 2. Key findings and recommendations from our study for hypotheses related to trait-based
693 differences in social network position

694

	Dyadic-based network data	**Grouping-event based network data**
Permutation-based approach	Low false positive rate when there is no assortativity Low false negative rate when there is positive or no assortativity. Slight deterioration in performance when network density is lower	High false positive rates, especially with positive assortativity or when larger numbers of grouping events are sampled Low-intermediate false negative rates. Deterioration in performance when there is negative assortativity or network density is lower
Exponential random graph models	Low false positive rate when there is negative or no assortativity Low false negative rate. Slight deterioration in performance when network density is lower or there is negative assortativity	High false positive rates especially with positive assortativity or when network density is higher Low false negative rate when there is no or positive assortativity and when network density is higher
Recommendations	**Both approaches generally perform well for dyadic-based network data** **We recommend that both approaches are viable for analysing dyadic-based network data, although ERGMs perform marginally better in most situations. We highlight the need for caution when confounding effects of assortativity are present until new methods are developed.**	**Standard ERGMs also suffer from high false positive rates and so do not present a ready-made alternative to datastream permutations to test network measure-trait relationships in grouping event-based networks (see Weiss et al 2020)** **We recommend careful use of node-label permutations (combined with appropriate correction for variation in sampling among individuals) until other methods have been evaluated for use on grouping event-based data**

695

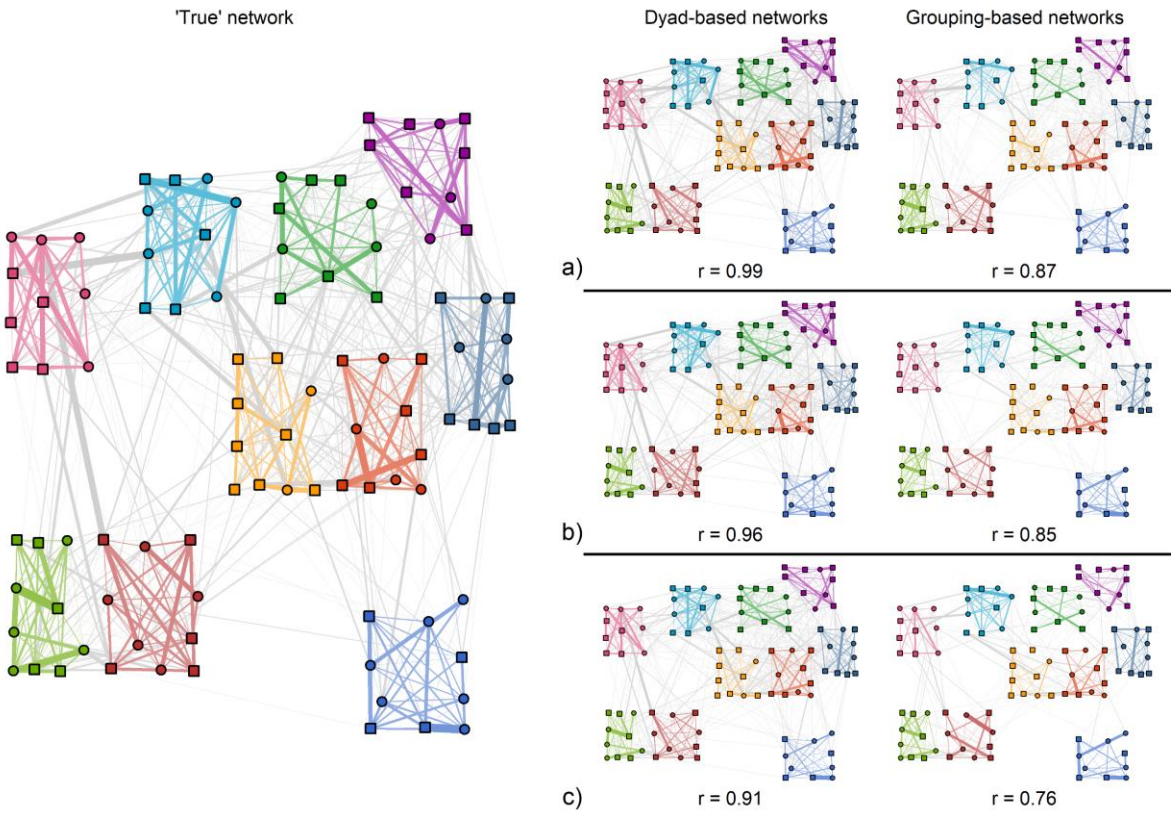
696

Figure legends

Figure 1. Example generated true generated network, alongside dyad-based and group-based networks at observation efforts of a) 0.9, b) 0.6 and c) 0.3. The results of mantel test comparisons between the dyad-based and group-based networks and the true network are presented underneath. Node colours represent the groups assigned at network generation. Round nodes are female while square nodes are male. Node position is approximately based on the spatial location of groups assigned at initial generation. Edge width indicates connection strength and edge colour whether a connectionis within a group (coloured as group) or between groups (black). Parameters used in generating this network were: *distance effect*= 4, *within-group edge density* = 0.8, *between-group edge density* = 0.4, *male effect*= 0 and *sex effect* = 1.

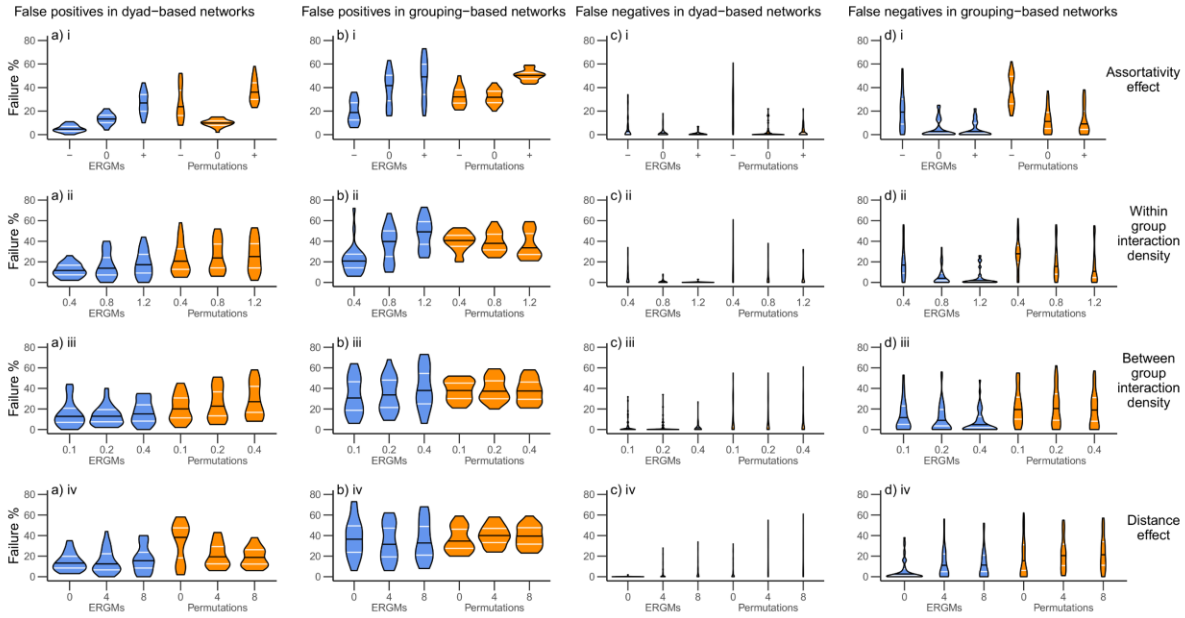
Figure 2. The failure rate per 100 simulations of the ERGMs (blue) and the permutation-based approach (orange) when detecting the difference between the sexes in strength. Row a) shows how the rates change due to the presence of negative, no, or positive assortativity by sex. Row b) shows how the rates change due to the strength of within-group interactions. Row c) shows how the rates change due to the strength of between-group interactions. Row d) shows how the rates change due to the strength of the effect of the distance between the groups. Plots show either the rate of false positives (columns i & ii), or the rate of false negatives (columns iii & iv), in both dyad-based networks (columns i & iii) and grouping event-based networks (columns ii & iv). The black bars indicate the medians, the white bars the 25% and 75% quartiles. The width of each violin relative to others within the plot gives the relative frequency of failure rates compared to other frequencies within that specific plot.

Figure 3. The failure rate of ERGMs (orange) and our permutation-based approach (blue) in dyad-based (a & b) or grouping event-based (c & d) networks at either correctly identifying the lack of effect (i.e. avoidance of false negatives; a & c), or correctly detecting the presence (i.e. avoidance of false positives; b & d) of the difference between the sexes in strength under a range of observation efforts. The black bars indicate the medians, the white bars the 25% and 75% quartiles. The width of each violin relative to others within the plot gives the relative frequency of failure rates compared to other frequencies within that specific plot.



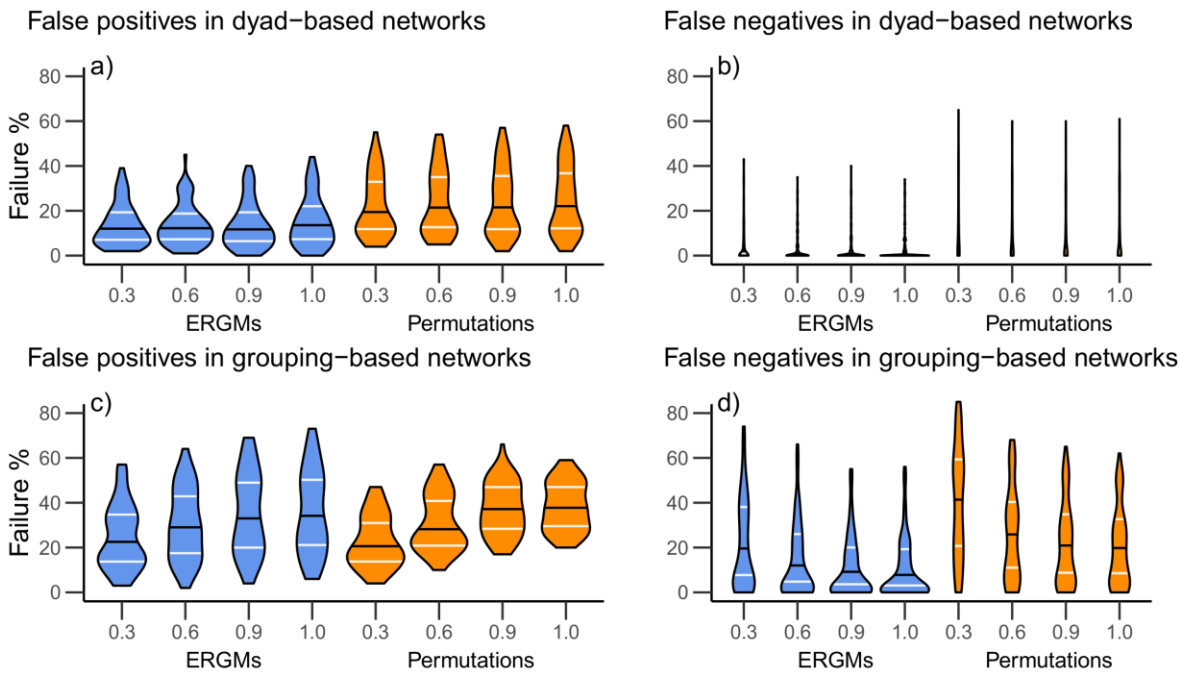
729

730 Figure 1



731

732 Figure 2



733

734 Figure 3